Genome and Symptoms based Disease Prediction

Mrunal Kurhade
Computer Engg. Dept, S.P.I.T
Mumbai, India
mrunal.kurhade@spit.ac.in

Meghan Lendhe
Computer Engg. Dept, S.P.I.T
Mumbai, India
meghan.lendhe@spit.ac.in

Shreya Patel
Computer Engg. Dept, S.P.I.T
Mumbai, India
shreya.patel@spit.ac.in

I. Literature Survey

Translation of genomic knowledge for use in medical practice is a highly anticipated goal. Disease causing genomes can be identified by various methods. Few methods are discussed in this paper. Thus, genetic-based predictive models can have profound impact in diagnosis and detection. However, given that (i) symptoms/disease are affected by genetic and environmental factors, (ii) the genetic view of susceptibility to diseases is not well-understood, and (iii) replicable susceptibility alleles, in combination, account for only a moderate amount of disease heritability, prediction of disease using genomes is difficult [8]. Considerable progress has been observed with extensive research(in progress) that identify disease causing genotypes. Some approaches of disease prediction have been published. We have summarized the present state of disease prediction, procedures followed for genome and symptoms based prediction. Disease prediction on the basis of symptoms and genes has now become a critical task. The traditional methods of clinical analysis of the diseases were time consuming and tedious. Thus machine learning can be used to analyze the data and diagnose the disorders accurately. Machine learning employs supervised and unsupervised learning techniques and mainly focuses on automatic learning from datasets [2]. Most of the disease prediction models use data mining and machine learning techniques [1]. Fuzzy Hierarchical Approach, Artificial Neural Networks (ANNs), Genetic algorithms (GAs), along with hybrid methods including a combination of some of these methods are used.

Deoxyribonucleic acid (DNA) and ribonucleic acid(RNA) are nucleic acids which form an essential macromolecule. The DNA has a double helix structure and carries all the genetic information. Each nucleotide is composed of nitrogen-containing nucleobase cytosine [C], guanine [G], adenine [A] and thymine [T]. Bioinformatics helps in extracting, analysing and classifying information from the gene sequence. ANN, despite being widely used for machine learning in bioinformatics, is only suitable for hierarchy structure not for flat [2]. Thus techniques like Support Vector Machine (SVM), Artificial Neural Network (ANN), Artificial NeuroFuzzy inference System (ANFIS), Naïve Bayes, etc have their own pros and cons. Thus correct combination of techniques must be chosen. A k-mer is a nucleotide sequence of a certain length(k). It typically refers to all the possible substrings of length k that are contained in a string [3]. k-mer can indicate low quality or contamination in sequences. Thus Pathway analysis is preferred to extract and explain the k-mer sequence based DNA sequence analysis [4].

The interacting co-conspirators resulting in common diseases are environmental and heritable components. Environmental components include chemical exposure, infection and caloric flux and heritable components include DNA variants, methylation patterns, and epigenetic RNA effects. The use of disease predisposing alleles and its discovery presents a challenge to the human genetics community. Parallel sequencing, RNA expression and high-throughput genotyping are the recent technological advances that have accelerated interrogation of genetic variation. Providing detailed and mechanistic insight into the molecular pathogenesis of disease states are some important uses of such discoveries. In determining robust clinical care, accurate prediction of potential disease can play a critical role for many diseases with effective treatments that may avert severe disease. It is essential to characterize the important aspects that produce useful predictive models.

A large number of genes will unlikely have considerably important predictive value over traditional risk factors if these variants predispose the risk factors. Intermediate outcomes like hypertension or dyslipidemia or smoking involves genes associated with cardiovascular disease [5] [6].

Most diseases are triggered by a combination of genomes. Gene to gene interaction affect the functionality of the genes. The generic factors of a disease can be identified by finding frequent sequences in diseased samples. Thus DNA can be used for prediction of disease. Disease can be predicted by identifying the disease causing genes present in DNA. Pathways are visualized for all the disease causing genes and the possibility(expressed in probability) of a disease is predicted. The probability is obtained by discarding redundant data and considering only those genes that have a strong effect [4].

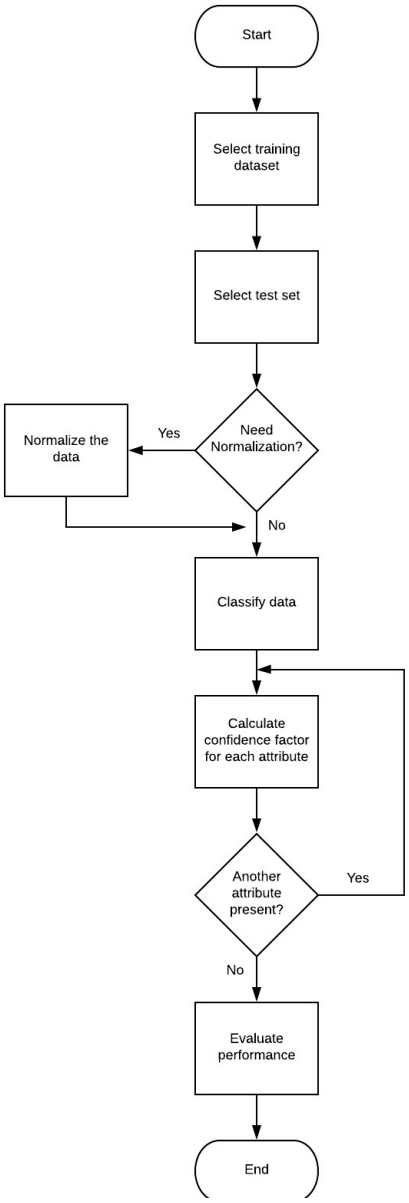
Interpretation and use of genomic knowledge in the healthcare is very challenging. At the moment, prediction of common diseases using genomes and symptoms is still not informative. Such prediction has been observed to vary a lot due to numerous factors (diet, age, BMI, etc.). Prediction of susceptibility to a disease is most difficult in individuals at early stages of the disease, but this category is highly important and prediction in this category is very meaningful. Strong effects can be seen from the interaction of a gene with other genetic variants (studied using pathways [4]). From a healthcare perspective, the current knowledge base is inadequate to

implement useful genome-based applications [7].

After studying some prediction methods we conclude that prediction using genomes and symptoms is weak and inaccurate even today. Current knowledge is insufficient to develop accurate disease prediction applications. But considering the pace at which research is going on, machine learning algorithms are being developed, analysis of genomes using big data,etc. the accuracy of present methods is bound to improve by a healthy margin.

II. Genomic Predictor for Diseases

Disease prediction based symptoms and genome analysis is one of the most interesting and challenging task. The lack of experts and wrongly analyzed cases has required the need to build up a quick and effective detection system. The main objective is to distinguish the key patterns or gene patterns from the medical data using the classifier model to precisely predict the disease.



1. Identifying Causal Genes and Diseases

When a clinician makes his or her diagnosis, they rely on a variety of information to support their decision: patient behaviors and symptoms, lab values, radiological tests, etc. The results ultimately get codified into formal classification systems, such as ICD-10. There may be follow-up tests, second and third opinions before a disease is finally confirmed. Rarely is there a single determining factor. The misalignment that commonly occurs between what the user wants to know and what available data and methods can actually tell us about the problem.

The impact of genetic-based predictive models on clinical decisions and therapy choice could be profound. However, given that (i) medical traits result from a complex interplay between genetic and environmental factors, (ii) the underlying genetic architectures for susceptibility to common diseases are not well-understood, and (iii) replicable susceptibility alleles, in combination, account for only a moderate amount of disease heritability, there are substantial challenges to constructing and implementing genetic risk prediction models with high utility. In spite of these challenges, concerted progress has continued in this area with an ongoing accumulation of studies that identify disease predisposing genotypes.

Multiple lines of evidence strongly support the notion that the large majority of common, chronic diseases have complex causes. Machine can predict diseases but cannot predict the sub types of the diseases caused by occurrence of one disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. The prediction system are broad and ambiguous.

First, the prevailing systems are dearer only rich people could pay for to such calculation systems. And also, when it comes to folks, it becomes even higher. Second, the guess systems are non-specific and indefinite so far. So that, a machine can envisage a positive disease but cannot expect the subtypes of the diseases and diseases caused by the existence of one bug.

Multiple predisposition variants may not yield perfect prediction of complex diseases alone. Thus question remains whether it will enhance the expectation of disease past conventional risk factors. Despite the fact that the development of profiles comprising of hereditary and ecological hazard factors shows up an obvious solution, studies so far demonstrated that hereditary components don't significantly enhance the expectation of type 2 diabetes and prostate cancer, yet again the number of genes examined was little.

2. Data Mining For Disease Prediction

Data Mining is a standout amongst the most imperative and motivating region of research with the goal of finding important data from huge data sets. The result of data mining innovations are to give benefits to healthcare association for gathering the patients having similar type of diseases or medical problems with the goal that healthcare association gives them effective treatments . Such techniques are also

used to analyze the different factors that are responsible for diseases. We are using the Classification approach to solve our problem. It is a supervised learning approach. It divides data samples into target classes and predicts them for each data point.

This concept is based on statistical learning theory. The SVM creates a hyper plane or multiple hyper planes in high dimensional space useful for classification. To maximise the separation between data points, SVM creates a hyper plane in original input space. To make this separation easier, original finite dimensional space is mapped into new higher dimensional space. For non-linear mapping of training samples, kernel functions are used. Thus the SVM works on the principle that data points are characterized using hyper plane which is constructed with help of support vectors. SVM is one of the approaches that are used by researcher in healthcare field for classification. Fei proposed Particle Swarm Optimization SVM (PSO-SVM) approach for analyzing arrhythmia cordis .

3. Calculating Confidence Factors

Our application will be at affordable cost. Calculating diseases as well as calculating other thinkable sub diseases. Broadly open by all at cheap cost. Data mining used (both supervised and unsupervised learning). For instance, a model may estimate the probability of an abnormal lab value. These method(s) will give probability of contracting a disease based on the genetic profile and symptoms. If they turn out to be accurate enough, they can be used for efficient and cost effective disease prediction. Genetic variations may enhance disease prediction past conventional risk factors when they are engaged with obscure pathways or in pathways with unmeasurable transitional factors. New yet obscure pathways might be more probable for a few diseases than for others.

III. Future Work

We have planned to develop a system that is integrated with a large database and will give the disease subtypes and the confidence factor as its output. This model has heavy usage of the machine learning algorithms. The model will be having a high processing power as it will have to handle a large dataset efficiently. Thus, the processor must be fast and the computational time should be minimum.We will retrieve the data from database, normalize it and apply the algorithm on it.

This approach consists six modules of data extraction, preprocessing , classification of genes, generating disease models, application of rules and finally prediction of diseases. The existing genome data is collected considering other attributes such as age, sex, etc . This collected data is given to the system. The system gives set of diseases associated with genes. Further result is given for calculating the confidence factor.

It will display the set of diseases and confidence factor to the user; the results will be highlighted if the confidence factor is greater than a threshold value. Given this model, it will serve as a great help to predict the diseases on the basis of genes and symptoms.

References

[1] Dhaval Raval, Dvijesh Bhatt, Malaram K Kumhar, Vishal Parikh, Daiwat Vyas, "Medical Diagnosis System Using Machine Learning," Volume 7 • Number 1 Sept 2015 - March 2016 pp.177-182.

[2] Pooja Dixit, Ghanshyam I. Prajapati, Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing, 2015 Fifth International Conference on Advanced Computing & Communication Technologies. pp: 41-47

[3] Mayank Pahadia, Akash Srivastava, Divyang Srivastava, Dr. Nagamma Patil, "Genome Data Analysis using MapReduce Paradigm," 2015 Second International Conference on Advances in Computing and Communication Engineering. pp: 556-559

[4] Syeeda Farah and Sushma M S, Dr.Asha T, Cauvery B and Shivanand K " DNA Based Disease Prediction using pathway Analysis," 2017 *IEEE* 7th International Advance Computing Conference pp: 629-634

[5] Kathiresan, S., Melander, O., Anevski, D., Guiducci, C., Burtt, N.P., Roos, C., Hirschhorn, J.N., Berglund, G., Hedblad, B., Groop, L, et al., " (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events,"N. Engl. J. Med., pp.358, 1240–1249

[6] Thorgeirsson, T.E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A. et al., "(2008) A variant associated with nicotine dependence, lung cancer and peripheral arterial disease,"Nature, pp.452, 638–642.

[7] A Cecile JW Janssens and Cornelia M van Duijn," Genome-based prediction of common diseases: methodological considerations for future research," Published: 18 February 2009 Genome Medicine 2009, 1:N (doi:10.1186/gm20). pp: r166-r173

[8] Steven J. Schrodi, Shubhabrata Mukherjee, Ying Shan, Gerard Tromp, John J. Sninsky, Amy P. Callear, Tonia C. Carter, Zhan Ye, Jonathan L. Haines, Murray H. Brilliant , Paul K. Crane, Diane T. Smelser, Robert C. Elston and Daniel E. Weeks," Genetic-based prediction of disease traits: prediction is very difficult, especially about the future," published: 02 June 2014 doi: 10.3389/fgene.2014.00162. pp:5-162

[9] Holtzman, N.A. and Marteau, T.M.,"Will genetics revolutionize medicine?," (2000) N. Engl. J. Med., pp.343, 141–144.

[10] Vineis, P., Schulte, P. and McMichael, A.J.," Misconceptions about the use of genetic tests in populations.," Genome Research, vol. 20, no. 9, pp. 1297–1303, 2010.

[11] A. McKenna," The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data," (2001) Lancet, pp.357, 709–712.

[12] Wray, N.R., Goddard, M.E. and Visscher, P.M.," Prediction of individual genetic risk to disease from genome-wide association studies,"(2007) Genome Res., pp.17, 1520–1528.

[13] Sebastian Okser, TapioPahikkala and TeroAittokallio," Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives,"BioData Mining, 6:5, 2013. pp: 1-16

[14] Li Ding, Michael C. Wendl, Daniel C. Koboldt and Elaine R. Mardis,"Analysis of next-generation genomic data in cancer: accomplishments and challenges,"Human Molecular Genetics, R1–R9, 2010. pp: 188-196

[15] Brand A, Brand H, Schulte in den Bäumen," The impact of genetics and genomics on public health,"*Eur J Hum Genet* 2008, 16. pp: 5-13

[16] Haga SB, Khoury MJ, Burke W,"Genomic profiling to promote a healthy lifestyle: not ready for prime time," *Nat Genet* 2003, 34:pp.347-350.

[17] Gunay, Adem Karadag,"Predicting Functional Regions in Genomic DNA Sequences Using Artificial Neural Network,"International Journal of Engineering Inventions Volume 3, Issue 6 (January 2014). pp: 2278-7461

[18] "Application of Data mining in Bioinformatics,"Indian journal of computer of science and engineering Volume 1 No.2 pp.114-118.

[19] MdSaiful Islam , Md MahmudulHasan, Xiaoyi Wang , Hayley D. Germack and MdNoor-E-Alam," Review ASystematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining,"Human Molecular Genetics, R1–R9, 2010.

[20] Mohammad Ahmad Alkhatib,Amir Talaei-Khoei,"Analysis of Research in Healthcare Data Analytics et al," Australasian Conference on Information Systems Al Khatib et al 2015, Sydney.

[21] Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh,“ Initial sequencing and analysis of the human genome,”Nature 2001, pp.409, 860–921.